Summer 6-11-2018

A Comparison of the Predictive Ability of Logistic Regression and Time Series Analysis on Business Credit Data

Lauren Staples
lstaple6@students.kennesaw.edu

A Comparison of the Predictive Ability of Logistic Regression and Time Series Analysis on Business Credit Data

Lauren L. Staples
Analytics and Data Science Institute
Kennesaw State University, Kennesaw, GA

Jennifer L. Priestley
Analytics and Data Science Institute
Kennesaw State University, Kennesaw, GA

Abstract— The credit industry creates models to determine the risk of lending money to consumers as well as to commercial customers. These models are heavily regulated in the U.S. as well as in other countries. Model inputs must be explainable to customers as well as to regulators. Two such modeling approaches that are currently commonly used are logistic regression models and time series models. This paper steps through the pre-processing and model building of these two models on a large commercial dataset and compares the predictive ability of these two methods. The two models achieved similar accuracy results: the logistic model had an accuracy of 89.6% while the time series model had an accuracy of 89.3%.

Keywords—Logistic, Time Series, Forecasting, ARIMAX.

I.INTRODUCTION

The credit industry profits when lending money to individuals or business accounts who pay loans back, but loses money for accounts that default on their loans. Credit lenders must predict (within laws and regulations) which applicants to lend money to and which to reject. While at first glance it may seem most profitable to have a high bar set for accepting applications, the truth is that a lot of "money is left on the table" for those accounts which were not accepted due to strict models, and those individuals or accounts that are rejected from one company will perhaps make money for competitors. In other words, reducing both false negatives (accounts predicted to not default that actually default) as well as false positives (accounts predicted to default that do not) are both opportunities for increasing revenue.

Logistic Regression Models are widely accepted in the credit industry by both regulatory agencies and credit lenders and have been used for decades (5). Time series models are also widely accepted. The goal of this paper is to compare the predictive ability of these two models in an applied setting. This paper describes the common methods of preparing the two different models from a base dataset that has been pre-processed specifically for the purposes of comparison.

II.BACKGROUND AND DEFINITIONS

The business problem presented by the credit industry lends itself well to a logistic regression model. A logistic model has a dependent variable that is a binary indicator of the class label for each case (in this paper, "0" will be the label for customers who do not default and "1" will be the label for customers who default). The logistic model takes a vector of input variables and determines the logit of the posterior probability of the dependent variable as a linear combination of those inputs (1). The logit is the log of the odds ratio, where the odds are the probability (P) of an event (default) happening vs the probability of an event not happening:

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad (4)$$

where X_1, X_2,…, X_k are predictor variables and the βs are the respective coefficients. The logistic model, in our case, predicts the probability of customer default based on input variables. The logistic model is performed on a dataset that does not necessarily have a time component. However, a "cross-section" (data taken at a constant time point) of a time series data set is a good candidate for a Logistic Model.

Another popular model for fitting and forecasting time series data is the autoregressive integrated moving-average model (ARIMA). This model was popularized by Box and Jenkins in the 1970s (SAS Institute Inc., 2014). This model is

appropriate for data that has observations over time at a specific time interval. It is especially appropriate for time series data that may have seasonality (vary cyclically over seasons). Time series models can take input variables as well. In our case, we will forecast the dependent variable for future values based on past values of the dependent variable as well as past and present values of the input variables. The general form of the equation this paper models is the model that includes inputs (ARIMAX, where the X stands for "transfer function"):

$$W_t = \mu + \sum_i \frac{\omega_i(B)}{\delta_i(B)} B^{k_i} X_{i,t} + \frac{\theta(B)}{\phi(B)} a_t \qquad (2)$$

where W_t is the target variable, μ is the mean term, $X_{i,t}$ is the *i*th input time series (or difference of the *i*th input series) at time t, $\phi(B)$ is the autoregressive operator, $\theta(B)$ is the moving average operator, and a_t is the random error. ARIMAX models are often described as orders of p, d, q which are parameters that further determine the autoregressive operator, the periodic differencing, and moving average operators (2).

III.DATA DISCOVERY

The data for this analysis is a de-identified set of business credit seeking applications provided by Equifax. Each observation is a business account, identified by unique Market Participant Identifiers (MPID). The data spans 8 years: from 2006 to 2014, in quarterly increments, and contains account history and payment status for utilities, business account, non-financial accounts and other variables that are typical information to what a lender would have access to on any business credit applicant. Each quarterly dataset contains approximately 11 million unique observations and 305 variables (28 of which are post-hoc information and therefore were immediately excluded, plus 4 irrelevant variables such as snapshot date were eliminated). Each quarterly dataset represents a cross-sectional slice of a time series dataset when considering all quarterly datasets stacked as one large dataset (referred to here as the "time series dataset"). The combined time series dataset thus has about 32 observations for each MPID (eight years of quarterly datasets is 32, unless some time periods are missing for an account). The cross-section of data selected to perform the logistic model is the second to last time point of July 31, 2014. All analyses in this paper were performed using SAS 9.4 software.

A. Assignment of Dependent Variable

In this analysis, we need to select a variable that lends itself well to both a Logistic and a Time Series Model and that can be thought of as a proxy for credit risk. The variable WSTNFpay3mon captures the worst non-financial payment status over the past three months. Since these are quarterly datasets, this captures the worst non-financial payment status for each business applicant in the quarter time interval. This variable represents how many billing cycles an account has been delinquent, if any. For the logistic model, this variable will be transformed into a "good/bad" class assignment, with a "0" being "good" and a "1" being "bad." Here, a "1" represents a defaulted business account. The logistic model target variable "goodbad" was transformed by selecting a conservative cutoff of 1 delinquent payment cycle. In other words, accounts with WSTNFpay3mon status less than or equal to 1 were transformed to zeroes in the logistic target variable "goodbad." For the time series model for which we compare predictive ability to the logistic model, the raw form of WSTNFpay3mon will be used for model fitting, but the variable will be transformed to a binary variable "goodbad," just like for the logistic model, in order to compute accuracy.

B. Missing Data and Imputation

All observations missing the dependent variable were deleted from the dataset. One fourth of the available MPIDs in the Time Series Dataset were selected to reduce the computational time of the time series analysis portion of the analysis. This means data surrounding 232,604 MPIDs entered both the logistic model as well as the time series model, but the number of observations for the logistic model were 232,604 unique MPIDs whereas about 7.4 million observations entered the time series model. Figure 1 below shows the relationship between the two datasets.

Figure 1: The Relationship Between the Logistic Dataset and the Time Series Dataset.

Logistic Dataset			Time Series Dataset	
MPID	Date		MPID	Date
G2200067K1103256788Y	31-Jul-14		G2200067K1103256788Y	31-Jan-06
G2200067K110395678X7	31-Jul-14		G2200067K1103256788Y	30-Apr-06
G2200067K110425678Y8	31-Jul-14		G2200067K1103256788Y	31-Jul-06
G2200067K1104Q5670V4	31-Jul-14		G2200067K1103256788Y	31-Oct-06
G2200067K11058567973	31-Jul-14		G2200067K1103256788Y	31-Jan-07
G2200067K11066567942	31-Jul-14		G2200067K1103256788Y	30-Apr-07
G2200067K1107456787W	31-Jul-14		G2200067K1103256788Y	31-Jul-07
G2200067K11077567891	31-Jul-14		G2200067K1103256788Y	31-Oct-07
G2200067K1107S567809	31-Jul-14		G2200067K1103256788Y	31-Jan-08
G2200067K1108O56701X	31-Jul-14		G2200067K1103256788Y	30-Apr-08
G2200067K11100567813	31-Jul-14		G2200067K1103256788Y	31-Jul-08
G2200067K11107567940	31-Jul-14		G2200067K1103256788Y	31-Oct-08
G2200067K1111656785U	31-Jul-14		G2200067K1103256788Y	31-Jan-09
G2200067K1111S567816	31-Jul-14		G2200067K1103256788Y	30-Apr-09
G2200067K11124567931	31-Jul-14		G2200067K1103256788Y	31-Jul-09
G2200067K111255679WZ	31-Jul-14		G2200067K1103256788Y	31-Oct-09
G2200067K1112S5678VZ	31-Jul-14		G2200067K1103256788Y	31-Jan-10
G2200067K1116O567823	31-Jul-14		G2200067K1103256788Y	30-Apr-10
G2200067K1118S5678XX	31-Jul-14		G2200067K1103256788Y	31-Jul-10
G2200067K1119T5678V0	31-Jul-14		G2200067K1103256788Y	31-Oct-10
G2200067K111RQ567889	31-Jul-14		G2200067K1103256788Y	31-Jan-11
G2200067K111S156704W	31-Jul-14		G2200067K1103256788Y	30-Apr-11
G2200067K111SQ56788W	31-Jul-14		G2200067K1103256788Y	31-Jul-11
G2200067K119075678X1	31-Jul-14		G2200067K1103256788Y	31-Oct-11
G2200067K119145679W8	31-Jul-14		G2200067K1103256788Y	31-Jan-12
G2200067K1192S567973	31-Jul-14		G2200067K1103256788Y	30-Apr-12
G2200067K1193R5678VU	31-Jul-14		G2200067K1103256788Y	31-Jul-12
G2200067K1194156784W	31-Jul-14		G2200067K1103256788Y	31-Oct-12
G2200067K1196Q56782U	31-Jul-14		G2200067K1103256788Y	31-Jan-13
G2200067K1196R567849	31-Jul-14		G2200067K1103256788Y	30-Apr-13
G2200067K119725678UV	31-Jul-14		G2200067K1103256788Y	31-Jul-13
G2200067K1198S567887	31-Jul-14		G2200067K1103256788Y	31-Oct-13
G2200067K1199P567840	31-Jul-14		G2200067K1103256788Y	31-Jan-14
G2200067K119NT5679TZ	31-Jul-14		G2200067K1103256788Y	31-Jul-14
G2200067K119PP567998	31-Jul-14		G2200067K1103256788Y	31-Oct-14

Table 1 below shows the proportion of data for each level of the dependent variable. The dependent variable is 75.2% 0s and 24.8% 1s. This is an acceptable split in that proportionally, it does not fall in the "rare event" category (where the event of interest occurs very infrequently in the dataset) (1). In the case for rare events, disproportionally over-representing the event cases is an additional step often necessary to achieve a functional model, but that is not the case here (1).

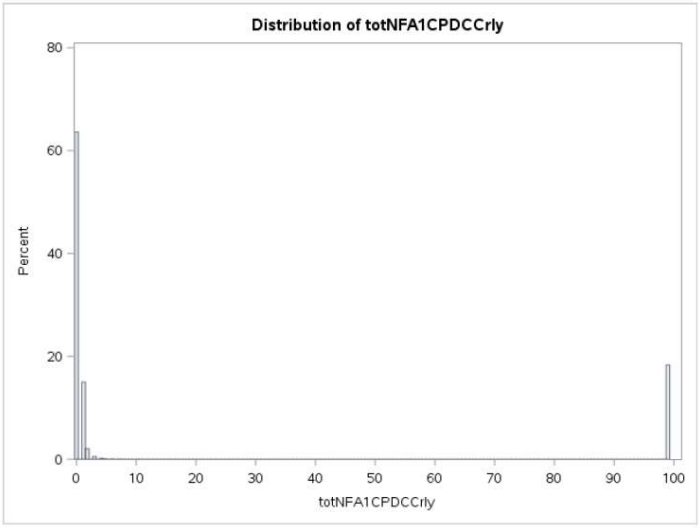
Table1: Frequency of the Dependent Variable.

The FREQ Procedure

goodbad	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	174853	75.17	174853	75.17
1	57751	24.83	232604	100.00

The dataset had many coded values that were evident as outliers in histograms for each variable. Closer inspection revealed that the coded values were typically "9" of some form: 9, 99, 999, 9999999, or even 9.999. Figure 2 shows a histogram for a variable containing a coded value of 99.

Figure 2: Histogram for Predictor Variable totNFA1CPDCCrly, which has a coded value of 99.



Coded values, in the interest of analysis, may as well be missing values. Both coded and missing values occur in the datasets and both need to be imputed. Imputing missing data is important because the logistic model operation will simply throw out observations that have missing variables. Imputing missing values in a dataset that we plan to compare two analysis methods means we have to select an imputation method that is appropriate for both methods. A univariate imputation method by median replacement is simple and while appropriate for a logistic model, since it uses the values of other observations (business accounts) to determine the median, it could introduce inappropriate variance in a Time Series Model, and perhaps confound seasonal trends. These coded values were thus replaced as missing for this two-step imputation process. Variables with greater than 30% missing values were deleted. A cutoff of 30% missing is an aggressive approach aimed at eliminating as many predictor variables as possible to make subsequent modeling activities more operational. The tradeoff is that one may exclude a meaningful variable by using such an aggressive approach. Judging by the predictive accuracy of your final model, one could always go back and include more variables by selecting a less aggressive cutoff for missing values. In our case: a total of 125 variables were eliminated due to having more than 30% missing or coded values.

Next, interpolating the missing values based on previous and following quarterly values for each account was investigated as a method for imputation. However, it is the case in this dataset that many accounts that are missing

data for one quarterly dataset are missing that data for all the quarterly datasets, and thus there is nothing to interpolate. It became reasonable then to assume that a missing value is really a zero, so the choice was made to then treat the coded and missing values as zeroes. There was no other missing data pattern found by visual inspection.

After imputation, a cross-sectional "slice" of the time series data set was taken at July 31, 2014, so that the two modeling methods have the same pre-processed data for which to compare predictive ability.

C. Variable Selection via Clustering

During imputation, the number of variables were reduced to 146 potential predictors (not including the MPID, date, WSTNFpay3mon or Goodbad). 146 potential predictors is still too many to be operational; in addition, we need to examine the variables for redundancy and multicollinearity. Multicollinearity in predictor variables presents many issues in modeling, one being that resulting models become sensitive to input data and small changes in data could result in large changes in the best-fitting model (1). It is best to pick one variable in a cluster of variables that are correlated, since correlated variables are also redundant. This removal of correlated variables both increases model simplicity, stability, and improves computational time. This dataset had three versions of several variables, a 3 month, 12 month and 24 month version. For example, the variable WstIpay3mon also had WstIpay12mon and WstIpay24mon. The WstIpay24mon variable stands for "worst industrial payment status over the past 24 months," and similarly for the 12 month and 3 month variable. Therefore, the 24 month variable will include the 12 and 3 month payment status if the 12 or 3 month worst payment status is also the worst over the past 24 months. Since the dependent variable WstNFpay3mon was selected, the 12 and 24 month versions of variables were discarded.

The cross-sectional dataset at July 31, 2014 was used for variable reduction by clustering. Clustering is an unsupervised concept, referring to the fact that it does not depend on the target/dependent variable (1). Variable clustering is closely related to a Principal Components Analysis (PCA), in that they are linear combinations of the original variables (1). However, an advantage of variable clustering over PCA is the interpretability of the coefficients, unlike PCA which retains nonzero coefficients for all eigenvectors, will only have nonzero coefficients between disjoint subsets of variables (uncorrelated clusters) (1). Having zero coefficients means those variables can be eliminated. Furthermore, using the variable clustering technique allows the flexibility of selecting a representative variable from each cluster instead of using the synthetic linear combinations as variables. This is an advantage in a regulated industry which requires traceability and explanation of why a model is rejecting or approving credit, as well as proof that discriminatory factors such as age or gender were not included.

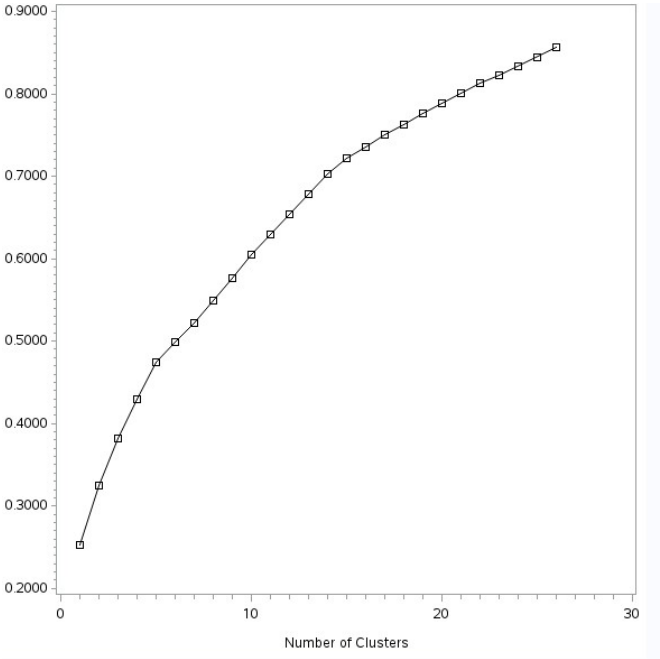
The choice was made to use the cross-sectional dataset for clustering due to computational time. All continuous variables entered the variable clustering procedure to find groups of variables that are correlated within the group but uncorrelated to those groups in other clusters. This concept is quantified by the 1-R^2 ratio, which is the 1-R^2 value within a variable's cluster over the 1-R^2 value of a variable to outside clusters (4). Each variable in each cluster has the 1-R2 ratio calculated for it in Table 1 below. A single variable is then selected as a representative variable from each cluster and carried forward for modeling. The variable with the lowest 1-R^2 is selected.

Table 1: Selecting Representative Variables from each Cluster.

Cluster	Variable	RSquare
Cluster 1	WstIpay3mon	0.5811
	totIA1CPDC3mon	0.7823
	totIA2CPDC3mon	0.2222
	totIA3CPDC3mon	0.2862
	totNFA3CPDC3mon	0.7466
Cluster 2	NoIAc3mon	0.2869
	NoIAcbalance3mon	0.5144
	NoIAccur	0.313
	NoNFA3mon	0.1218
	NoNFAbalance3mon	0.3739
	NoNFAcur	0.1814
	NoOpenNFA224	0.2158
	NoSasNFA	0.2036
	NoSasNFA3mon	0.1939
Cluster 3	BrtInd	0.0009
	JudInd	0.0034
	LienInd	0.0031
	LienJudInd	0.0004

While variable reduction is achieved by variable clustering, there is a tradeoff of information loss. This information is captured in Figure 3 below. If all the variables were included, the total variation would be 100%. However, the variable clustering procedure on our data found that 26 clusters explained 86% of the variation.

Figure 3: Proportion of Variation Explained by Clusters.



One further evaluation of the variable clustering on our data is to examine the variance inflation factors (VIF's). Linear regression was performed on the logistic model dataset (by using the original dependent variable Wstnfpay3mon) to calculate the VIFs. VIFs are indicators of multicollinearity among variables, and the equation is below.

$$VIF = \frac{1}{(1-R^2)} \quad (6)$$

The continuous variables retained in the logistic dataset after variable clustering were run through a linear regression. All VIF's were below the industry-accepted cutoff of 10 (as seen in Figure 4).

Figure 4: Variance Inflation Factors for Numeric Variables Remaining after Variable Clustering.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.64055	0.01124	145.93	<.0001	0
totIA2CPDC3mon	1	1.01457	0.00742	136.67	<.0001	5.61480
NoNFA3mon	1	-0.00264	0.00085251	-3.09	0.0020	3.08799
LienJudInd	1	-0.02602	0.00065894	-39.49	<.0001	1.05515
NoTA3mon	1	0.04256	0.00147	28.95	<.0001	1.90077
totNFPD	1	-0.00000498	1.286168E-7	-38.73	<.0001	1.56554
totNFA4CPDCCrly	1	0.69610	0.00859	81.08	<.0001	2.11482
totNFA1CPDCCrly	1	0.17782	0.00391	45.46	<.0001	3.35550
YearsinBusiness	1	0.00009491	0.00008858	1.07	0.2840	1.05721
Industry	1	-0.01137	0.00080635	-14.10	<.0001	1.11282
totLAllLiens	1	5.909839E-8	1.085384E-8	5.44	<.0001	1.01492
NoNFChgAcc3mon	1	4.51715	0.02632	171.63	<.0001	1.13435
pctNFPDAmtst3mon	1	-0.00062672	0.00000528	-118.65	<.0001	1.09714
HstIB3mon	1	-4.42629E-8	7.494049E-9	-5.91	<.0001	1.05177
NoEmployeeRange	1	0.02629	0.00156	16.86	<.0001	1.19820
totNFA1CPD3mon	1	0.05595	0.00365	15.34	<.0001	1.96984
totNFA3CPD3mon	1	0.05455	0.00826	6.60	<.0001	1.94547
NFA3monCurRate	1	-0.86356	0.00669	-129.02	<.0001	1.27251
totC3NFPDAmt3mon	1	-0.00000422	3.862605E-7	-10.91	<.0001	1.18336
totLAllJud	1	2.552482E-8	1.776051E-8	1.44	0.1507	1.00519
pctSasNFA	1	-0.00029397	0.00001244	-23.62	<.0001	1.04751
NoClosedNFA226	1	-0.04484	0.00285	-15.72	<.0001	1.26065
totIA2CPD3mon	1	-0.79681	0.01061	-75.09	<.0001	3.13165
NAICSCode	1	-5.96032E-8	8.87324E-9	-6.72	<.0001	1.15173
NoNewNFAcc3mon26	1	-0.01770	0.00406	-4.36	<.0001	1.24318
totC2NFPDAmt3mon	1	1.102218E-7	4.510017E-7	0.24	0.8069	1.29946

After variable clustering, a total of 24 numeric variables remained as potential model inputs.

D. Logistic Modeling

Discretizing and transforming input predictor variables can often simplify or even improve supervised models. Predictors are rarely optimized in their original form. Discretization can take many different forms, such as user-defined equal width or by equal frequencies. These discretized variables can then be transformed using odds and log odds. These methods were not used in this paper, the reason being that while discretizing and transforming variables may be appropriate for the logistic model, it may not be appropriate for a direct performance comparison with the time series model.

The cross-sectional dataset at July 31, 2014 was then split into a training set and a test set at 70% and 30%, respectively. The time series dataset had to be split by keeping the time series values for each MPID found in either the logistic training set or logistic test set, respectively.

The logistic regression procedure was run on the logistic training set and scored by the test set. A backward-selection stepwise model was run with all remaining 24 numerical predictor variables (the 11 categorical variables were not entered). A backwards-selection stepwise model was selected to ensure every variable was considered by the modeling function. The alternative, a forward-selection method, is start-point sensitive (1).

The ARIMAX approach was used to fit several models, with the best model being selected by minimizing the AIC (Akaike information criterion). Unlike the logistic procedure which has a stepwise function to test many models, the ARIMAX was run many times manually. While seasonality was suspected, using a differencing term of the intervals resulted in a much higher AIC than ARIMA models with only one differencing term (necessary due to nonstationarity observed in the data). The AIC did not vary much as different terms were added or removed manually, however including at least some input variables was important as this improved the AIC over models depending on autoregression with WSTNFpay3mon alone.

IV.RESULTS

For the logistic model, the prediction on the test dataset are probabilities ranging from 0 to 1. Multiple cutoffs were examined for determining the predicted "goodbad" variable, an accuracy of 89.6% was found on the logistic test data set. Figure 5 shows the confusion matrix for this model and one can see the false positives and false negatives are balanced. False positives occur when a customer is predicted to default when in fact they would not have defaulted, and a false negative is when a customer is predicted to not default when in fact they will defaulted. The credit-lending company would lose money in both cases; however often a false positive is more costly than a false negative. Profit can be optimized by appropriately tuning (balancing) the rate of false positives and false negatives according to the profit model of the company.

Figure 5: Confusion Matrix for Logistic Model on Test Data.

Table of goodbad by preds			
	preds		
goodbad	0	1	Total
0	55707 72.57 96.54 90.30	1995 2.60 3.46 13.24	57702 75.17
1	5986 7.80 31.41 9.70	13072 17.03 68.59 86.76	19058 24.83
Total	61693 80.37	15067 19.63	76760 100.00

Figure 6: C-Statistic for Logistic Model on the Test Data.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.4	Somers' D	0.788
Percent Discordant	10.6	Gamma	0.788
Percent Tied	0.0	Tau-a	0.294
Pairs	4532923643	c	0.894

The variables kept in the model were used as candidate input variables in the time series model (see Appendix). A time point of July 31, 2014 was once again used to select a cross-section of the test time series data for accuracy that is comparable to the accuracy found with the logistic model. The forecasted value of WSTNFpay3mon was transformed to a "goodbadpred" variable by using the same cutoff as finding the "goodbad" variable from the ground truth WSTNFpay3mon, of less than or equal to 1 being "good." The accuracy was found to be 89.3%, consistent with the accuracy found by logistic regression. One can see, however, that the false positives and false negatives are not as well balanced as in the Logistic Model. However, the Time Series Model does a better job at reducing false negatives but at the expense of false positives. Tuning the cutoff point for the forecasted values might improve the balancing, but may have a tradeoff in accuracy. Figure 7 below shows the confusion matrix for the time series model.

Figure 7: Confusion Matrix for the Time Series Model.

Table of goodbad by goodbadpred			
	goodbadpred		
goodbad	0	1	Total
0	50482 65.77 87.49 98.07	7220 9.41 12.51 28.55	57702 75.17
1	991 1.29 5.20 1.93	18067 23.54 94.80 71.45	19058 24.83
Total	51473 67.06	25287 32.94	76760 100.00

V. CONCLUSION

In conclusion, the Logistic Model and the Time Series Model produce similar overall accuracies: 89.6% to 89.3%. However, the Time Series Model depends on the Logistic Model for determining its input variables and is much more computationally expensive than the Logistic Model alone. However, the insight into the missing value pattern that was provided by an examination of the time series dataset (that the coded values were actual zeros) improved the accuracy of the logistic model from 84.9% (where observations with missing values were simply discarded) to 89.6%.

VI. ACKNOWLEDGEMENTS

Thank you to Bob Vanderhyden and Edwin Baidoo for their help and discussion for this paper.

VII. REFERENCES

1. Potts, William J.E, and Michael J Patetta. Predictive Modeling Using Logistic Regression: Course Notes. SAS Institute Inc., 2008.
2. SAS/ETS® 13.2 User's Guide: The ARIMA Procedure. SAS Institute Inc., 2014.
3. SAS® Forecasting and Econometrics Community. SAS Forecast Studio - Scoring Existing ARIMAX Models on New Data (Independent Variables). Retrieved 4/29/2018 from https://communities.sas.com/t5/SAS-Forecasting-and-Econometrics/SAS-Forecast-Studio-Scoring-Existing-ARIMAX-Models-on-New-Data/td-p/136399
4. Rudd, Jessica M. MPH, GStat and Priestley, Jennifer L., "A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Non-Financial Payment Status in Commercial Credit" (2017). Grey Literature from PhD Candidates. 5. http://digitalcommons.kennesaw.edu/dataphdgreylit/5
5. Dixon, P. a. (2014, April 2). *U.S. Federal Trade Commission Public Comment Documents.* Retrieved from https://www.ftc.gov/system/files/documents/public_comments/2014/04/00007-89171.pdf
6. Variance Inflation Factors. National Institute of Standards and Technology, Statistical Engineering Division. https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/vif.htm. Accessed 5/30/2018.

VIII. APPENDIX

Variables used as inputs into the final logistic model:

Figure 8: Variables Included in Final Logistic Model and Their Contributions.

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
NoNFA3mon	1.0000	1.130	1.122	1.137
LienJudInd	1.0000	0.956	0.951	0.961
NoTA3mon	1.0000	1.044	1.032	1.056
totNFA4CPDCCrly	1.0000	2.023	1.892	2.162
totNFA1CPDCCrly	1.0000	6.466	6.211	6.731
YearsinBusiness	1.0000	0.999	0.998	0.999
Industry	1.0000	0.986	0.980	0.992
totLAllLiens	1.0000	1.000	1.000	1.000
NoNFChgAcc3mon	1.0000	12.468	8.225	18.901
pctNFPDAmtst3mon	1.0000	0.997	0.997	0.998
totNFA1CPD3mon	1.0000	0.324	0.313	0.335
NFA3monCurRate	1.0000	0.270	0.256	0.284
totC3NFPDAmt3mon	1.0000	1.000	1.000	1.000
pctSasNFA	1.0000	1.000	1.000	1.000
NoClosedNFA226	1.0000	0.753	0.734	0.773
NoNewNFAcc3mon26	1.0000	1.059	1.024	1.095
totC2NFPDAmt3mon	1.0000	1.002	1.002	1.002

Variables used as input into the final Time Series Model:

Figure 9: Variables used in the ARIMAX Model and Their Contributions.

Variable NoClosedNFA226 has been differenced.

Correlation of wstnfpay3mon and NoClosedNFA226	
Period(s) of Differencing	1
Variance of input =	0.148304
Number of Observations	39999
Observation(s) eliminated by differencing	1

Variable pctSasNFA has been differenced.

Correlation of wstnfpay3mon and pctSasNFA	
Period(s) of Differencing	1
Variance of input =	70774.18
Number of Observations	39999
Observation(s) eliminated by differencing	1

Variable NoNFChgAcc3mon has been differenced.

Correlation of wstnfpay3mon and NoNFChgAcc3mon	
Period(s) of Differencing	1
Variance of input =	0.0099
Number of Observations	39999
Observation(s) eliminated by differencing	1

Variable totC2NFPDAmt3mon has been differenced.

Correlation of wstnfpay3mon and totC2NFPDAmt3mon	
Period(s) of Differencing	1
Variance of input =	9179278
Number of Observations	39999
Observation(s) eliminated by differencing	1

Variable NoTA3mon has been differenced.

Correlation of wstnfpay3mon and NoTA3mon	
Period(s) of Differencing	1
Variance of input =	1.067102
Number of Observations	39999
Observation(s) eliminated by differencing	1

Variable totNFA1CPD3mon has been differenced.

Correlation of wstnfpay3mon and totNFA1CPD3mon	
Period(s) of Differencing	1
Variance of input =	0.461862
Number of Observations	39999
Observation(s) eliminated by differencing	1